**Chapter 5 lecture questions**

**Q1:** A $k$ nearest neighbour model is used to predict summer temperature (warm, normal or cool) from the spring climate conditions $\mathbf{x}$. With $k = 10$, for a test point $\mathbf{x}'$, the 10 nearest neighbours from the training data have 3 warm, 3 normal and 4 cool summers. What is the posterior probability of the coming summer being (a) warm, (b) normal and (c) cool?

**Answer:** (a) 3/10, (b) 3/10, (c) 4/10.
Solution:
The probability of $y'$ being in class $C_1$ is $n_1/k$, ..., in class $C_i$ is $n_i/k$, so the probability of being warm is 3/10, normal is 3/10 and cool is 4/10. Classification involves choosing the class with the highest probability, so predict "cool" for the summer.

---

**Q2:** Compute the 2×2 table of $P(Y|X)$ for the tornado problem.

**Solution:**
Given the table for $P(X, Y)$ and $P(X)$ :

| | $P(X,Y)$ | | |
|---|---|---|---|
| $Y$: | $X$: L | NL | $P(Y)$ |
| T | 20/100 | 5/100 | 25/100 |
| NT | 10/100 | 65/100 | 75/100 |
| $P(X)$ | 30/100 | 70/100 | |

From

$$P(X, Y) = P(Y|X)P(X), \tag{1}$$

we get

$$P(Y|X) = P(X, Y)/P(X). \tag{2}$$

Use this formula and the elements from the table for $P(X, Y)$ and $P(X)$ to get:

| | $P(Y\,\vert X)$ | |
|---|---|---|
| $Y$: | $X$: L | NL |
| T | 20/30 | 5/70 |
| NT | 10/30 | 65/70 |
| Sum | 30/30 | 70/70 |

---

**Q3:** Suppose a test for a toxin in a lake gives the following results: If a lake has the toxin, the test returns a positive result 99% of the time. If a lake does not have the toxin, the test still returns a positive result 2% of the time. Suppose only 5% of the lakes contain the toxin. What is the probability that a positive test result for a lake turns out to be a false positive?

**Answer:** 27.7%.

Solution:

The two classes for $Y$ are $C_1$ and $C_0$, being toxic and non-toxic, respectively. The measurements $X$ are either 1 (test result positive) or 0 (result negative). We are given that $P(C_i)$ has the values $P(C_1) = 0.05$ and $P(C_0) = 1 - 0.05 = 0.95$. For $P(X|C_i)$, we are given $P(1|C_1) = 0.99$, and $P(1|C_0) = 0.02$. We want to find $P(C_0|X = 1)$, the probability there is no toxin yet the test came out positive. From Bayes' theorem, we have

$$\begin{aligned} P(C_0|1) &= \frac{P(1|C_0)P(C_0)}{P(1|C_0)P(C_0) + P(1|C_1)P(C_1)} \\ &= \frac{0.02 \times 0.95}{0.02 \times 0.95 + 0.99 \times 0.05} = 0.277. \end{aligned}$$

Hence the probability for a false positive is a surprisingly high 27.7%.

---

**Q4:** Hierarchical clustering analysis (using Ward's method) was applied to a dataset containing the concentration of 11 ions in the honey from 44 locations in Europe (Fermo et al., 2013). How many clusters for the locations appear optimal?

**Solution:**

In the dendrogram figure, after 2 clusters (top red line), the distances between clusters become smaller and similar in size for 3-6 clusters. After 6 clusters (bottom red line), the distances become even smaller for 7 clusters or more. Seems reasonable to pick either 2 clusters or 6 clusters as optimal.

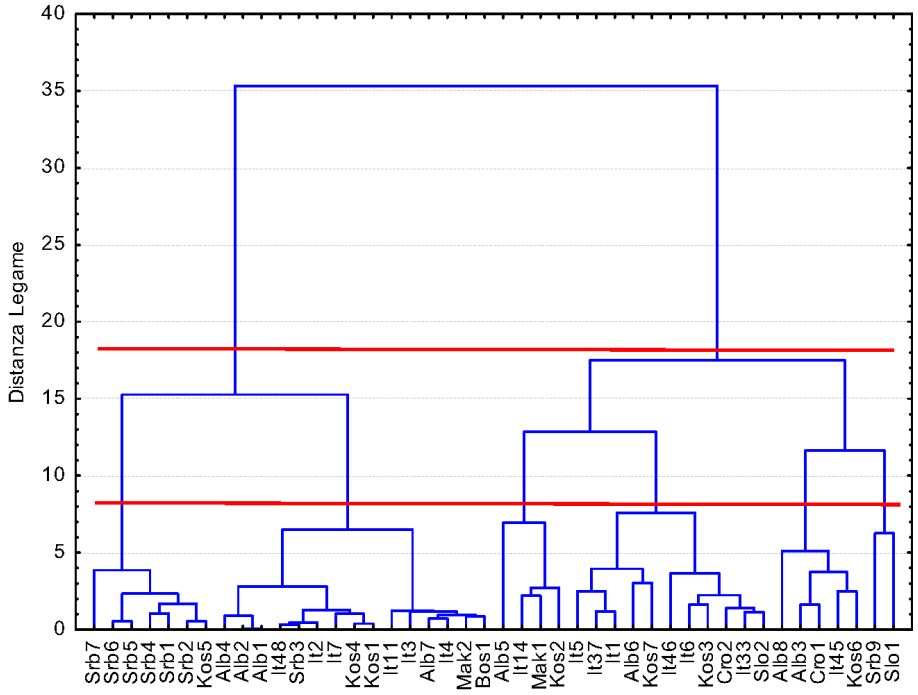


**Fig. 4.** Dendrogram obtained from PCA (Ward's method) of honey samples from Italy and from W. Balkans.

# References

Fermo, P., Beretta, G., Facino, R. M., Gelmini, F., and Piazzalunga, A. (2013). Ionic profile of honey as a potential indicator of botanical origin and global environmental pollution. *Environmental Pollution*, 178:173–81.